

ASWIN JOSE

Lead ML Engineer · Applied AI / LLM Systems · Bengaluru, India
mathew.aswin.jose@gmail.com · linkedin.com/in/aswin-jose · github.com/meagoodboy

SUMMARY

ML engineer who owns systems end to end, from research and training through deployment and the customer conversation. Owns the ML audio-analysis engine behind a health-AI startup's live product, alongside the company's backend and app engineers: rebuilt the core algorithm from scratch into the pipeline now running in production, shipped a multi-step LLM pipeline with human-in-the-loop review for clinical reporting, and led ML for a BMW pilot that proved vitals extraction in a previously untested in-cabin setting. Four publications in ML for healthcare. B.Tech CS + MS, IIIT Hyderabad.

EXPERIENCE

Lead ML Engineer, Sonus Health / Decorte Future Industries

Mar 2025 – Present
London, UK (Remote)

Health AI · smartphone-based cardiac screening

- Own the ML audio-analysis engine behind the live product, alongside the company's backend and app engineers: problem framing, data, modelling, deployment, and monitoring.
- Rebuilt the core algorithm from the ground up: migrated the legacy PCG pipeline to a new architecture, then built a dedicated veterinary heart-sound pipeline for murmur detection from standard smartphone microphones, which now powers the live Sonus Health product.
- Shipped a production multi-step LLM pipeline for clinical report generation: chains sequential LLM calls over model output and clinician input with human-in-the-loop review, producing final reports veterinarians actually use.
- Led ML engineering for a BMW pilot deploying the vitals-extraction engine in-cabin via seatbelt-mounted microphones, hitting >80% accuracy across key vitals and proving viability in a previously untested use case. Worked directly with BMW stakeholders.
- Built the production reliability layer: neural pipeline paired with a tabular classifier, probability calibration, and tiered confidence bucketing so the product only surfaces results it can stand behind. First-authored two 2026 arXiv papers on the system.

Software Engineer, Couture.ai

Jun 2024 – Mar 2025
Bengaluru, India

Personalisation & search at e-commerce scale

- Built and optimised a RAG-based ML search pipeline for a large-scale e-commerce platform (1M DAU), improving retrieval quality and processing efficiency.
- Designed an asynchronous, high-throughput service for smart querying over large product catalogs; applied model-optimisation techniques to cut inference latency in production.
- Deployed and tuned the pipeline hands-on with customer engineering teams against their data, traffic patterns, and business requirements.

Undergraduate Researcher, CCNSB (IIIT Hyderabad)

May 2021 – Jun 2024
Hyderabad, India

Representation learning for cancer genomics

- Representation learning on multi-omics cancer data with graph auto-encoders and VAEs, benchmarked against state-of-the-art; built an end-to-end patient survival-prediction pipeline and adapted it to per-cancer-type subtasks.
- Network-diffusion approach to renal cell cancer stratification, identifying biologically relevant subtypes and the genetic factors driving survival. Published (journal + preprint).

Machine Learning Intern, VoltaML

Jul 2022 – Oct 2022
California, US (Remote)

Inference optimisation

- Cut XGBoost inference time using the llaives compiler, outperforming state-of-the-art approaches like treelite by at least 2.5×; benchmarked regression and diffusion models for inference speed and deployability.

TECHNICAL SKILLS

ML / AI PyTorch, LLM pipelines & prompt orchestration, RAG, embeddings & vector search, model calibration, GNNs, XGBoost, audio & signal ML, ONNX

Engineering Python, C++, C, JavaScript, Shell, SQL, FastAPI, Docker, Kubernetes, AWS, Kafka, Redis, PostgreSQL, MongoDB, Qdrant, Git

EDUCATION

B.Tech in Computer Science + MS in Computational Natural Sciences, IIIT Hyderabad

2019 – 2024

Winner, Institute Research Award (AY 2022–23) for publishing a journal article during undergraduate study.

PUBLICATIONS

- Sonus Health: Calibrated Heart-Murmur Detection from Smartphone-Based Veterinary Auscultation. Jose A, Decorte RPJE, Locquet L. arXiv, 2026. arxiv.org/abs/2606.24885
- Instant Preliminary Cardiac Analysis from Smartphone Auscultation: A Real-World Canine Heart Sound Dataset and Evaluation. Jose A, Decorte RPJE, Locquet L. arXiv, 2026. arxiv.org/abs/2601.13593
- DeepGraphMut: A graph-based deep learning method for cancer prognosis using somatic mutation profile. Jose A, Srivastava A, Vinod PK. bioRxiv, 2024. doi.org/10.1101/2024.12.03.626568
- Network diffusion-based approach for survival prediction and biomarker identification in papillary renal cell carcinoma. Shetty KS, Jose A, Bani M, Vinod PK. Mol Genet Genomics 298, 871–882, 2023. doi.org/10.1007/s00438-023-02022-4